Handling Fungal data in MoBeDAC

Jason Stajich
UC Riverside
Fungal Taxonomy and naming undergoing a revolution

One fungus, one name

Chair: Pedro Crous
CBS Fungal Biodiversity Centre, Utrecht, Netherlands
Vice-chair: Keith Sofert, Agriculture and Agri-food Canada, Ottawa, Canada

Fungi have far higher diversity than land plants. There could be more than two million species but this estimate is very tentative because fungal taxonomy is so incomplete (only 50,000 species have been formally described).

Although no fungal species are known to be endangered, members of this kingdom are important as disease agents, as food, as producers of antibiotics, as agents of fermentation and as the basis of much organic decay.

Past taxonomic work on fungi has been slowed by their morphological conservatism and by the small size of most species. Thus, DNA-based taxonomy will revolutionize our understanding of fungal diversity and enable, for the first time, the connection of their life stages. BOL will register barcodes for at least 10,000 fungal species by 2014 with a particular focus on building barcode libraries for indoor fungi, for basidiomycetes (the "higher fungi") and for those fungi that are important pathogens of agriculture and forestry.

Goal
10,000 species

Also posted in Barcode Library, WG 1.3 - Fungi, Working Group Profile
http://www.biology.duke.edu/fungi/mycolab/primers.htm
http://www.biology.duke.edu/fungi/mycolab/primers.htm
Why is this hard?

• ITS is easier to amplify because it is multicopy and concerted evolution keeps the copies homogenized(*) and rRNA genes will change more slowly helping make universal primers possible(*)

• Curated sequence database of marker to taxonomy needs to be built

• Taxonomy specified to different depths for some lineages, especially early branching ones

• ITS cannot really be used to build trees - it is a good barcoding molecule as it changes rapidly. Though in some lineages not rapidly enough and resolving in those lineages requires another marker

• LSU is good for backbone and major grouping but hard to resolve species or even genus often with this molecule.

* Assumptions that mostly/often hold true
Speaking the same language

- Unified Taxonomy
- Multiple marker sequences
  - ITS, SSU, LSU
  - COI1
- Assembling the Fungal Tree of Life markers
  - (RPB1, RPB2, EF1alpha)
- Phylogenomically chosen 40-60 protein coding genes.
What’s in GenBank for ITS? (ca 2010)

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neocallimastigomycota</td>
<td>267605</td>
</tr>
<tr>
<td>Ascomycota</td>
<td>79597</td>
</tr>
<tr>
<td>Basidiomycota</td>
<td>45393</td>
</tr>
<tr>
<td>Glomeromycota</td>
<td>3809</td>
</tr>
<tr>
<td>Chytridiomycota</td>
<td>485</td>
</tr>
<tr>
<td>Microsporidia</td>
<td>482</td>
</tr>
<tr>
<td>Blastocladiomycota</td>
<td>10</td>
</tr>
</tbody>
</table>

1-454 run
Some ITS databases

- UNITE (unite.ut.ee) - UNITE barcoding sequences: 3878 ITS sequences of 1508 species from 255 genera
  Fungal ITS sequences in database (UNITE + INSD): 205,688

- Extracts from GenBank ~

- In case of uncurated data there are many mis-specified taxonomy assignments to sequence.
  - Worst are endophytic fungi that are assigned to plants or other improper specimen identification

- Need: Expert curation, collection. Much of this is being done very well by UNITE team and collaborators. We (Sloan funded project) do want to help
LSU & databases

- Ribosomal Database Project (MSU) has a Naive Bayesian classifier for Fungal LSU - [http://rdp.cme.msu.edu/classifier/](http://rdp.cme.msu.edu/classifier/)

- Still new to us to be able to test out (Liu et al 2011) but promising and quite fast as it doesn’t have same alignment-based

- Vast majority of data being generated seems to be ITS, but sometimes there is a paired LSU study.
Playing with real data

- Amend et al PNAS 2010 “Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics.”

- 72 samples of fungi from 6 continents. Sampled ITS2 region and the D1-D2 region of LSU with 454-FLX

- Main finding of increasing species diversity with increasing latitude
Fig 1. Amend et al 2010
## MG-RAST with Fungal Data

### Technical

**Anthony Amend (UC Berkeley)**  
*Plant and Microbial Biology, United States of America*

### Metagenomes

![Export Jobs Table](image)

<table>
<thead>
<tr>
<th>MG-RAST ID</th>
<th>Metagenome Name</th>
<th>Size (bp)</th>
<th>Biome</th>
<th>Location</th>
<th>Country</th>
<th>Sequence Type</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>4484996.3</td>
<td>Rgl00.ITS</td>
<td>2,916,194</td>
<td>dry; building</td>
<td>-</td>
<td>Canada</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4484997.3</td>
<td>SW01.ITS</td>
<td>870,516</td>
<td>dust</td>
<td>-</td>
<td>United States of America</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4484999.3</td>
<td>TTIUY.ITS</td>
<td>2,308,713</td>
<td>city; building</td>
<td>-</td>
<td>Uruguay</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485000.3</td>
<td>TTN4.ITS</td>
<td>1,113,476</td>
<td>dry; building</td>
<td>-</td>
<td>Netherlands</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485001.3</td>
<td>VM1.ITS</td>
<td>2,345,057</td>
<td>city; building</td>
<td>-</td>
<td>United States of America</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485002.3</td>
<td>VW01.ITS</td>
<td>1,217,526</td>
<td>dry; building</td>
<td>-</td>
<td>United States of America</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485003.3</td>
<td>WL1.ITS</td>
<td>2,440,816</td>
<td>dry; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485004.3</td>
<td>WL2.ITS</td>
<td>1,272,024</td>
<td>city; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485005.3</td>
<td>WL3.ITS</td>
<td>3,025,402</td>
<td>dry; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485006.3</td>
<td>WL4.ITS</td>
<td>2,121,128</td>
<td>city; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485007.3</td>
<td>WL5.ITS</td>
<td>2,345,057</td>
<td>dry; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485008.3</td>
<td>WL6.ITS</td>
<td>162,078</td>
<td>dry; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485009.3</td>
<td>WL7.ITS</td>
<td>1,899,453</td>
<td>city; building</td>
<td>-</td>
<td>Micronesia</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
<tr>
<td>4485010.3</td>
<td>z071193.ITS</td>
<td>538,101</td>
<td>dry; building</td>
<td>-</td>
<td>Uruguay</td>
<td>Amplicon</td>
<td><a href="#">submit</a> <a href="#">metadata</a> <a href="#">analysis</a></td>
</tr>
</tbody>
</table>
PROJECT INFORMATION

This dataset is part of a project studying fungi from temperate to tropical buildings.

Sequencing against ITS and 28S rRNA databases for each sample. View paper here: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928887

There are 127 other metagenomes in this project.

GSC MiX S INFO

<table>
<thead>
<tr>
<th>Investigation Type</th>
<th>Metagenome: Amplitcon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Name</td>
<td>Indoor fungi from temperate to tropical buildings</td>
</tr>
<tr>
<td>Latitude and Longitude</td>
<td>50.45417, -104.60667</td>
</tr>
<tr>
<td>Country and/or Sea, Location</td>
<td>Canada</td>
</tr>
<tr>
<td>Collection Date</td>
<td>2010</td>
</tr>
<tr>
<td>Environment (Biome)</td>
<td>Urban</td>
</tr>
<tr>
<td>Environment (Feature)</td>
<td>Building</td>
</tr>
</tbody>
</table>

SOURCE HITS DISTRIBUTION

6,148 (98.1%) of reads had similarity to ribosomal RNA genes.

The graph below displays the number of features in this dataset that were annotated by the different databases below. Those include databases with functional hierarchy information, and ribosomal RNA databases. The bars representing annotated reads are colored databases have different numbers of hits, but can also have different types of annotation data.

Download chart data

<table>
<thead>
<tr>
<th>ribosomal RNA genes</th>
<th>ITS</th>
<th>18S (1.9%)</th>
<th>16S</th>
<th>5.8 S (3.2%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Hits</td>
<td>506</td>
<td>473</td>
<td>43</td>
<td>10</td>
</tr>
</tbody>
</table>

Note: Sequences containing multiple predicted features are only counted in one category. Currently downloading of sequences via chart above is not enabled.
Hits summarized by different taxonomic levels
Rarefaction curve (1 sample)
Summary of Indoor Fungal Metagenomes using MG-RAST tools

Analyses performed with KbaseKit R package
( Kevin Keegan, Daniel Braithwaite)
R package to download and analyze MG-RAST annotated data

Install the KbaseKit

```r
> install.packages("KbaseKit.tar", type="source", repo=NULL)
> library(KbaseKit)
```
Batch download data from MG-RAST

```r
> my_data <- kbGet("4441679.3;4441680.3;4441682.3;4441695.3;4441696.3;4440463.3;4440464.3", "abundance", namespace="SEED", param="format/plain")
```

Save in simple format for R, Matlab, Excel etc.

```r
> write.table(my_data, file = "my_data_file", col.names=NA, row.names = TRUE, sep="\t", quote=FALSE)
```

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Nostoc azollae'</td>
<td>0708</td>
<td>mgid4441679.3</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abietitrichia deficiens ATCC 49176</td>
<td>236</td>
<td>172</td>
<td>263</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>50</td>
</tr>
<tr>
<td>Acaryochloris marina MBIC11017</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Acetobacter cellulolyticus CD2</td>
<td>118</td>
<td>101</td>
<td>124</td>
<td>5</td>
<td>5</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>Acetobacter pasteurianus IFO 3283-01</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acetathalibium arabadicum DSM 5501</td>
<td>31</td>
<td>20</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Acholeplasma laidlawii PG-8A</td>
<td>36</td>
<td>27</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Achromobacter plechaudii ATCC 43553</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>48</td>
<td>65</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Achromobacter mgidylesolezumidians AR</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>68</td>
<td>69</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Acidaminococcus fermentans DSM 20731</td>
<td>181</td>
<td>135</td>
<td>179</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Acidaminococcus sp. D21</td>
<td>86</td>
<td>80</td>
<td>121</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Acidilobus secalivorans 345-15</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acidimicrobium ferrooxidans DSM 10331</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acidiphilium cryptum JF-5</td>
<td>14</td>
<td>8</td>
<td>16</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Acidiphilium multivorans</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Preprocess (normalize and standardize the data)

> preprocessing(file_in = "my_data_file", produce_fig = TRUE)

Distribution of species level taxonomic abundances for 128 samples

After normalization and standardization, data are more comparable, but non-normal
PCA of normalized counts – Painted by rRNA type
Investigated ITS and 28S samples to determine taxa that exhibit the most significant differences

Abundance
Red(low) -> Green(high)

10% most abundant taxa that are significantly different between 28S and ITS
(Mann-Whitney test – Bonferroni adjusted p-value 0.05)
PCA of normalized counts – Painted by sampled country

unknown  red
Australia   blue
Canada     green
Indonesia  magenta
Mexico     yellow
Micronesia  cyan
Netherlands  raspberry
South Africa  orange
United Kingdom  spring_greer
United States  turquoise
Uruguay    ocean
VAMPS with Fungal data

- Testing the use of GAST and the UNITE ITS database on the Amend et al data.

- Good recall for this dataset - 8% of data is unknown, but still evaluating correctness of assigned taxa.

- Have also tested leave-one-out cross validation with test ITS data and there is reasonable ability to recall taxa.

- With MBL team, be testing an integration of ITS data into the standard VAMPS analyses.
Thanks

**UCR**
Steven Ahrendt
Daniel Borcherding
Raghu Ramamurthy
(FungiDB)

**VAMPS-MBL**
Sue Huse
Anna Shipunova
Mitch Sogin

**QIIME**
Gail Ackerman
Jesse Stombaugh
Rob Knight

**MG-RAST**
Daniel Braithwaite
Travis Harrison
Kevin Keegan
Andreas Wilke

Wednesday, March 7, 12